# `bionn` team solution preliminary write-up

Romuald A. Janik* and Michael A. Olesik†

Institute of Theoretical Physics
Jagiellonian University
Kraków, Poland.

## Abstract

We describe the second place solution in the Algonauts 2021 Challenge *How the Human Brain Makes Sense of a World in Motion* on predicting fMRI brain activations of human subjects watching movie clips. The Deep Neural Network features are projected to a varying number of receptive fields and aggregated over time, by taking the maximum in both cases. We observe a passage from local to more global features within the visual hierarchy. We introduced also a set of motion related data, extracted from the movie clips.

## 1 Introduction

The Algonauts challenge in 2021 [1], comprises the prediction of fMRI activations of 10 subjects while watching 1102 movie clips of 3s duration. The training data consists of 1000 clips shown with 3 repetitions for training the predictive models, while the predictions were evaluated by the organizers on the activations of the remaining 102 clips, which were shown with 10 repetitions.

The competition involved the predictions of (reliable) voxels from the V1, V2, V3, V4, EBA, FFA, LOC, PPA and STS ROIs within the mini-track, and a selected part of the whole brain within the full-track.

The key motivation for us for entering the competition was on the one hand, checking how such contemporary Deep Learning approaches as contrastive learning or transformers would fare in comparison to the more traditional Deep Learning classification models, and on the other hand to explore the intuitions about the relation of the visual hierarchy with the structure of Deep Neural Networks.

In both of these questions, the results were not entirely in line with our prior expectations. An additional intriguing question was how the dynamic time-dependent character of the movie clips would get reflected in the fMRI activations.

## 2 Key ingredients

We primarily adopted a straightforward pipeline consisting of

1. Evaluating neural network features for the subsampled individual frames
2. Coarse-graining the features into appropriate receptive fields
3. Accumulating the features across all the frames
4. Fitting a Ridge regressor on the resulting features and the *means* of the fMRI activations coming from the 3 repetitions

*e-mail: romuald.janik@gmail.com
†e-mail: olesik.michael@gmail.com

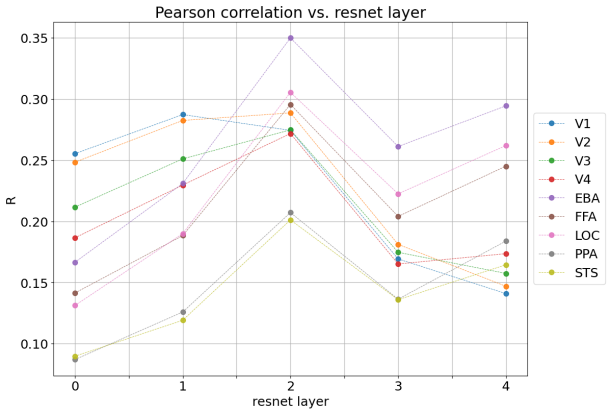


Figure 1: Cross-validated R correlation score for subsequent layers of `resnet152` for V1, V4, LOC, EBA, FFA, PPA & STS and `resnet50` for V2 and V3. The lines serve to guide the eye.

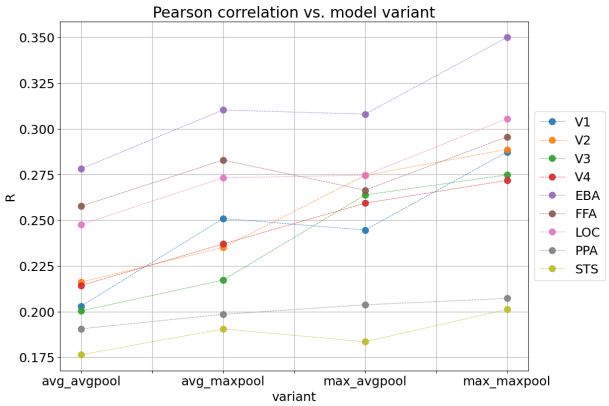


Figure 2: Cross-validated R correlation score for all ROIs calculated using variants with taking average or maximum over frames (`avg` or `max`) and over each receptive field (marked by `avgpool` or `maxpool`). The best performance is obtained by taking the maximum in both cases.

Additionally, the features were optionally augmented by separately estimated observables related to overall movement between the frames. Below we discuss in turn each of the above ingredients.

**Cross-validation procedure.** We adapted 5-fold cross-validation for each subject. The movie clips were shuffled differently for the individual subjects in order not to introduce an overall bias. The models were scored by the average $R$ over all subjects and folds.

**Neural networks.** We tested the classical `alexnet`, `vgg19`, and various ResNets as well two types of contrastive learning networks `pclv2` [2] and `simclr` [3] and an implementation of a visual transformer network [4].

The outcome was that the best CV results for all ROIs were obtained by the two largest ResNet networks: `resnet50` (V2[1] and V3) and `resnet152` (all other ROIs). In fact, the latter network was also quite close to the best scores also for V2 and V3, so it could be considered as the best overall.

[1]In this case an intermediate layer of `pclv2` was marginally better, but we did not use it.

The more modern neural network approaches and in particular the final contrastive features unfortunately did not give competitive results.

**Neural network layers.** For the ResNet networks, we extracted features from 4 intermediate layers (for the sizes of the layers look up Table 1) and the final classification layer[2]. Using CV we identified the layer which gives the highest R score for a particular ROI. Surprisingly enough, the same layer (`layer2`) provided the best explanation of the brain activations for *all* ROIs apart from V1, for which an earlier layer (`layer1`) was better (see Fig. 1).

| | |
|---|---|
| `layer0` | (256x56x56) |
| `layer1` | (512x28x28) |
| `layer2` | (1024x14x14) |
| `layer3` | (2048x7x7) |
| `layer4` | (1000) |

Table 1: Layer sizes of employed ResNet networks.

The ROIs located higher in the visual hierarchy did not benefit by going deeper into the neural network but rather by accumulating features into larger receptive fields as we discuss now.

[2]To this end, we used some of the code from the previous Algonauts challenge [6].

**Receptive fields.** The solution of one of the authors [5] to the previous Algonauts challenge [6] indicated that it is beneficial to reduce the resolution of the original neural network features using a set of "receptive fields" by taking the maximum over the receptive fields[3], reducing thus the effective spatial resolution of the features to $7 \times 7$, $5 \times 5$, $3 \times 3$ or even $2 \times 2$. We observed a clear correlation of the optimal resolution with the location of the ROI in the visual hierarchy ($7 \times 7$ for V1, V2 and V3, $5 \times 5$ for V4, $3 \times 3$ for EBA, FFA, LOC, PPA and $2 \times 2$ for STS).

We compared two methods of evaluating the features for a given receptive field: taking the *maximum* or the *average* (see Fig. 2). We found that the first method gives consistently better results, in agreement with the findings in [5].

**Accumulating features across frames.** The short duration of the clips and the relatively long time-scale of the fMRI response suggests that one should somehow integrate the features from all movie frames and only use that for prediction. This could be done by averaging over frames, as was done in the Starter Kit provided by the organizers. We found, however, that a better strategy was to take the *maximum* over the time frames (see Fig. 2).

**Machine learning regression model.** We fit the resulting accumulated features to the *mean activation* of the given voxels using a simple Ridge regressor with the regularization parameter $\alpha = 1e4$. We experimented with some nonlinear approaches but did not observe a benefit, probably due to the very high dimensionality of the data.

# 3 The baseline solution

The elements of the baseline solution for each ROI are summarized in Table 2. In all cases we take the maximum over the receptive fields and the maximum over movie frames. The test score of this solution **0.6446** would already be in the $2^{nd}$ place slot of the competition.

[3]Using the standard PyTorch `adaptive_max_pool2d`.

| ROI | network | layer | res. | test score |
|---|---|---|---|---|
| V1 | `resnet152` | `layer1` | $7 \times 7$ | 0.6497 |
| V2 | `resnet50` | `layer2` | $7 \times 7$ | 0.6564 |
| V3 | `resnet50` | `layer2` | $7 \times 7$ | 0.6495 |
| V4 | `resnet152` | `layer2` | $5 \times 5$ | 0.6822 |
| LOC | `resnet152` | `layer2` | $3 \times 3$ | 0.6872 |
| EBA | `resnet152` | `layer2` | $3 \times 3$ | 0.7070 |
| FFA | `resnet152` | `layer2` | $3 \times 3$ | 0.7065 |
| PPA | `resnet152` | `layer2` | $3 \times 3$ | 0.5731 |
| STS | `resnet152` | `layer2` | $2 \times 2$ | 0.4900 |
| OVERALL | | | | **0.6446** |

Table 2: Details of the baseline solution for the individual ROIs.

In the following we describe additional improvements on top of this baseline solution.

# 4 Incorporation of movement

The overall approach outlined in section 2, does not take into account the dynamic character of the movie clips. In particular the constructed features would not distinguish between an essentially static image and a clip with lots of motion.

One option to incorporate some of this information would be e.g. to add the difference of extracted neural network features between the last and first frame. This did not help, however, probably due to the very high dimensionality of the features in comparison to overall number of movie clips. We decided therefore to introduce features which were sensitive purely to movement irrespective of the semantic content of the movie clips.

We first estimated the vectors of frame by frame displacements by finding the shift in pixels in the range $(-15, 15) \times (-15, 15)$ which minimizes the mean square distance between one frame and the shifted succeeding frame. We also repeated similar analysis for 9 square patches of each frame. We then took the mean displacement over all frames obtaining $2 + 2 \times 9 = 20$ features. We also added the absolute values of those mean displacements to allow for sensitivity of the brain activations to the overall magnitude of the motion irrespective of direction.

| ROI | $\Delta_{score}$ |
|---|---|
| V1 | 0.0022 |
| V2 | unknown |
| V3 | 0.0008 |
| V4 | **0.0040** |
| LOC | 0.0007 |
| EBA | 0.0033 |
| FFA | 0.0002 |
| PPA | **0.0089** |
| STS | -0.0010 |

We show alongside the difference between the test scores of models with the features from the baseline solution with movement added and the original baseline scores.

We observe that the movement features give a small but noticable improvement for V4 (somewhat unexpectedly) and for PPA. For the final solution for these two ROIs we therefore add the movement features, leaving the rest for the moment unmodified.

# 5 Refining the receptive fields

Since the reduction of the higher resolution neural network features by taking the maximum over some receptive fields was very beneficial for the scores of the higher visual areas, we decided to investigate whether changing the sizes and/or overlaps of the receptive fields with respect to those given by the PyTorch `adaptive_max_pool2d` would better reflect the brain activations. Indeed, we observed benefits for the EBA, FFA, LOC and STS ROIs.

The `resnet152`'s `layer2` features have a spatial resolution of $14 \times 14$. The baseline receptive fields for the STS ROI were thus

$$[0,6]\times[0,6],\ [0,6]\times[7,13],\ [7,13]\times[0,6],\ [7,13]\times[7,13]$$

which can be understood as a cartesian product of the intervals $\{[0,6],\ [7,13]\}$. We found that adding an additional large central receptive field $[2,11] \times [2,11]$ was beneficial. But, surprisingly enough, one could then replace the previous 4 squares by the overall frame. So for STS we finally take

$$[0,13] \times [0,13] \text{ and } [2,11] \times [2,11]$$

This essentially means that STS responds best to completely global features.

For EBA, FFA and LOC, which in the baseline solution had $3 \times 3 = 9$ receptive fields, we also adjoin the large central receptive field $[2,11] \times [2,11]$. In addition, we make the original $3 \times 3$ receptive fields larger and overlapping. They are given by a cartesian product of

$$\{[0,5],\ [3,10],\ [8,13]\}$$

The central receptive field in the cartesian product could possibly be made slightly larger and the additional one eliminated, but due to lack of time, we did not test that.

| receptive field | EBA | FFA | LOC | STS |
|---|---|---|---|---|
| standard | 0.7070 | 0.7065 | 0.6872 | 0.4900 |
| modified | 0.7129 | 0.7161 | 0.6950 | 0.4930 |

Table 3: Comparison of test scores for the standard and modified variants of receptive fields.

In Table 3 we compare the test scores for the modified receptive fields with the original baseline ones.

# 6 The final solution

In Table 4 we present our final solution. It improves somewhat on the baseline one but still remains at the $2^{nd}$ place.

| ROI | network | layer | res. | test score |
|---|---|---|---|---|
| V1 | `resnet152` | `layer1` | $7 \times 7$ | 0.6497 |
| V2 | `resnet50` | `layer2` | $7 \times 7$ | 0.6564 |
| V3 | `resnet50` | `layer2` | $7 \times 7$ | 0.6495 |
| V4 | `resnet152` | `layer2`✓ | $5 \times 5$ | 0.6862 |
| LOC | `resnet152` | `layer2` | mod. | 0.6950 |
| EBA | `resnet152` | `layer2` | mod. | 0.7129 |
| FFA | `resnet152` | `layer2` | mod. | 0.7161 |
| PPA | `resnet152` | `layer2`✓ | $3 \times 3$ | 0.5820 |
| STS | `resnet152` | `layer2` | mod. | 0.4930 |
| OVERALL | | | | **0.6490** |

Table 4: Details of the final solution for the individual ROIs. The modified receptive fields are marked by *mod.* and are described in section 5. A checkmark ✓ indicates that the movement features described in section 4 were added.

## 7 The whole brain solution

For the whole brain solution, we did not pursue an independent construction but rather reused the best models developed for the individual ROIs. Since, as we saw, the features entering the best models for the various ROIs were different, we may expect that for the voxels in the whole brain a single model would not suffice, and one should choose a specific model for a particular voxel. We do not expect that concatenating the features from all these models would work well due to the large dimensionality, but we did not have time to verify whether this is indeed the case.

In order to associate a given model prediction to a particular voxel, we performed cross-validation on the whole brain data for all the models used for the individual ROIs, but now computing the R's for the *individual voxels*. Then for the whole brain prediction we used the prediction of the model with the highest CV score for the particular voxel. For the models from the baseline solution this procedure gives **0.3422** test score for the whole brain, while for those from the final solution we get **0.3483**.

We noticed, however, that the individual voxel CV scores may be quite noisy and not being a very good predictor of the optimal model. This problem appears when trying to apply the same procedure to a wider range of models as good individual voxel CV scores may also appear purely by chance. As an alternative, for the final whole brain solution, we overrode the choice from voxel CV scores for voxels belonging to one of the ROIs from the Mini-Track. In that case we used the model chosen for the particular ROI. We obtained a small test score increase to **0.3490**. This suggests that some spatial "smoothing" of the model choice for the individual voxels might be beneficial and worth investigating. All the whole brain solution described here have test scores in the $2^{nd}$ place slot in the competition.

## 8 Discussion

Our main observations in the process of working on solutions to the competition are as follows. The biggest variants of `resnet` networks outperform others with `resnet152` being the best for the majority of ROIs. The more recent constructions involving transformers and contrastive learning unfortunately seem to be further removed from the human visual system. Despite expectations, we do not observe an increase in the optimal layer depth[4] correlated with the place of the ROI in the hierarchy of the visual system. Indeed the intermediate `layer2` is optimal for 8 out of 9 ROIs (see Fig. 1). The main differences between the ROIs lies in aggregating features within different receptive fields, going to more global ones for the higher-level ROIs. The aggregation is optimally performed by taking the maximum rather than averaging (see Fig. 2), indicating that the brain acts primarily within the *winner-take-all* paradigm. A similar conclusion holds for aggregating over time. Moderate improvements can be obtained by adding movement related features (primarily for PPA and to a lesser extent V4), and enlarging the receptive fields with some overlap between the neighbouring ones for the higher level ROIs. The STS ROI seems most sensitive to almost completely global features.

**Acknowledgements.** We thank the organizers for an interesting competition. This work was supported by the Foundation for Polish Science (FNP) project *Bio-inspired Artificial Neural Networks* POIR.04.04.00-00-14DE/18-00.

## References

[1] R.M. Cichy, K. Dwivedi, B. Lahner, A. Lascelles, P. Iamshchinina, M. Graumann, A. Andonian, N.A.R. Murty, K. Kay, G. Roig, A. Oliva, *The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion*, arXiv:2104.13714

[2] Junnan Li, Pan Zhou, Caiming Xiong, Steven Hoi (2020), *Prototypical Contrastive Learning of Unsupervised Representations*, arXiv: 2005.04966

[4] This statement is valid on a coarse-grained level looking just at the main stages of the `resnet152` network. We did not check what happens within a single stage.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton (2020), *A Simple Framework for Contrastive Learning of Visual Representations* , arXiv: 2002.05709

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv: 2010.11929

[5] R.A. Janik, *Explaining the Human Visual Brain Challenge 2019 – receptive fields and surrogate features*, arXiv:1907.00950

[6] R. M. Cichy, G. Roig, A. Andonian, K. Dwivedi, B. Lahner, A. Lascelles, Y. Mohsenzadeh, K. Ramakrishnan, A. Oliva, *The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence.* arXiv: 1905.05675